Expanding Acceptable Transfer Requirements
Transfer Instructions for Permanent Electronic Records

WEB CONTENT RECORDS

1.0 PREFACE

As part of the Electronic Records Management (ERM) E-Gov Initiative to improve electronic records management in Federal agencies, and in cooperation with other Federal agencies, NARA is issuing guidance to supplement current requirements in 36 CFR 1228.270 for transferring permanent electronic records to NARA. Through the record scheduling process, NARA, in consultation with agencies, determines which *web content records*[1] are permanent and when transfer should occur.

This guidance expands currently acceptable formats to enable the transfer of permanent web content records to NARA.

2.0 INTRODUCTION

Web content records are a priority electronic records format identified by NARA and partner agencies as part of the Electronic Records Management (ERM) initiative, one of the twenty-four E-Gov initiatives under the President's Management Agenda. A major goal of this initiative is to provide the tools for agencies to access electronic records for as long as required and to transfer permanent electronic records to NARA for preservation and future use by Government and citizens.

In Fiscal Years 2002, 2003, and 2004, NARA worked with several Federal agencies to develop and promulgate transfer guidance for five formats: (1) e-mail with attachments, (2) scanned images of textual records, (3) records in Portable Document Format (PDF), (4) digital photographic records, and (5) digital geospatial data records. For more information on NARA's E-Gov ERM Initiative and the completed products, please visit our web site at: http://www.archives.gov/records_management/initiatives/erm_overview.html.

3.0 OPERATING PRINCIPLES

The web is evolving from many efforts, ideas, protocols, standards and formats. The following operating principles were used to establish specific transfer requirements in this guidance.

3.1 This transfer guidance is based on the assumption that the web content records being transferred have been scheduled by the agencies and appraised as permanent by NARA.

3.2 The *World Wide Web Consortium* (W3C) has provided the basic guidance for what the web is today and will be in the foreseeable future.

---

[1] Terms in italics are further defined as they apply to this document in the glossary in Appendix A

3.3 A *domain name* defines the administrative boundaries and content of an agency's web site unless a formal web management agreement specifically allows agency content to reside on a non-agency domain. In which case, content hosted at the non-agency domain is also included as part of the agency's web content.

3.4 Web content is limited to what is accessed over the *Hypertext Transfer Protocol* (HTTP[2]). All other transfer protocols (i.e., File Transfer Protocol and Simple Mail Transfer Protocol) are excluded.

3.5 Hypertext is a defining characteristic of the web; its functionality needs to be preserved.

3.6 *Hypertext Markup Language* (HTML[3]) is the predominant format in which web content records are viewed on a browser. Content in other formats such as TIFF or PDF are either embedded in the HTML or referenced by it.

3.7 The *client/server architecture* that characterizes the web involves complex server environments that are impractical to preserve.

3.8 A web page is what is sent from a server to a client *browser* when a *Uniform Resource Locator* (URL) has been activated and may include multiple image, text, audio or video files.

3.9 *Web content* presently is rendered in two forms - static and dynamic.

3.9.1 Static web content consists of information in the form of "web documents" that are rendered identically each time they are accessed.

3.9.2 Dynamic web content consists of information that is rendered differently based on specific user input and is usually managed in a database associated with a server.

4.0 Scope

This guidance applies to web content records managed by an agency that have been appraised and scheduled for permanent retention by NARA. It does not address content associated with *Web Services*, the *Semantic Web* or the *Wireless Web*.

4.1 Inclusions

Web content records that may be transferred to NARA under this guidance include:

[2] When HTTP is mentioned assume *Secure Hypertext Transfer Protocol* (HTTPS) as well.
[3] When HTML is mentioned assume XHTML as well.

4.1.1 Web content records sent via *Hypertext Transfer Protocol* (HTTP) from a server to a client *browser* when a *URL* has been activated.

4.1.2 Web content records that share a domain name including content managed under formal agreement and residing on another site.

4.1.3 All component parts of web content records that have been appraised as permanent including image, audio, video and all other proprietary formats.

4.1.4 Static and dynamic content.

4.2 Exclusions

The following will not be accepted for transfer under this guidance:

4.2.1 Program or administrative records documenting the management of web sites.[4]

4.2.2 Referenced content (e.g., accessed via hyperlink) that resides in a different domain and is not managed for an agency under a formal agreement.

4.2.3 Static images, such as screen shots, of web content records, because they do not retain hypertext functionality. (Note: PDF is not considered an image format).

5.0 EFFECTIVE DATE

The requirements in this guidance are effective September 17, 2004.

6.0 TRANSFER REQUIREMENTS FOR WEB CONTENT RECORDS

6.1 Web content records that are appraised as permanent must be scheduled for permanent retention on a Standard Form 115, Request for Records Disposition Authority (SF 115). The records must be organized as either a logical grouping of information or by agency records series.

6.2 NARA understands that legacy records and records whose disposition is changed from temporary to permanent may present unique circumstances for agencies. Any agency having permanent web content records that do not

[4] These records should be scheduled via submission of a Standard Form 115, Request for Records Disposition Authority (SF 115), or by applying General Records Schedule (GRS) 24, Information Technology Operations and Management Records.

meet the requirements in this guidance should contact the NARA appraisal archivist assigned to that agency (see section 8.0).

6.3 Web Content Criteria

The following sections specify the minimum technical requirements for transfers of web content records to NARA.

6.3.1 NARA will accept the following:

6.3.1.1 Source document(s) from the server in an open standard markup language (HTML or *Extensible Markup Language* (*XML*)).

6.3.1.2 Web content that can be read by a commonly available web browser.

6.3.2 Agencies must convert dynamically rendered content such as Cold Fusion (.cfm) files, Hypertext Preprocessor (.php) files, Server Side Inclusions (.shtml), and Active Server Page (.asp) files to an HTML version that can be read by a standard web browser.

6.3.3 Transfers must include the complete web content record. Agencies should use a method or combination of methods that captures the complete record. (See Appendix B)

6.3.3.1 When HTML or XML files are transferred, agencies must include all associated files that are pulled together to create the web content record being transferred. For example, image files will be transferred with the HTML file in which they are embedded.

6.3.3.2 Agencies must transfer Cascading Style Sheets used in HTML files transferred directly from the web server.

6.3.3.3 For XML files transferred from the web server, agencies must also transfer either an *XML schema* (WXS) or a *Document Type Definition* (DTD) and appropriate *style sheet*(s).

6.3.4 When collecting web content manually, agencies must transfer only *file fragments*, images, XML schema, DTDs, or style sheets associated with documents identified as part of the web content record. Files that are not part of the web content record but are collocated in a directory with pertinent files must not be included in the transfer.

6.3.5 Transfers of permanent web content may include image, audio, video and proprietary formats that are part of the complete web content record in their native format and as they appear on the web.

6.3.6 Agencies must retain scripts embedded in HTML documents (e.g., javascript) as part of the file as written.

6.3.7 Agencies must redirect all links within the web content record that is being transferred so that they remain active in the transferred records.

6.3.8 Agencies must disable all external links from the transferred web content records. (See Appendix B)

6.3.9 Agencies should include comments or external documentation describing the content of external links that they consider significant to the content of the web site being transferred. Comments should include the title of linked content, the URL, the source agency or institution, and the significance of the content to the web content records being transferred.

6.3.10 Agencies must transfer interactive content for database-driven, *server-side* dynamic environments within their transfers according to the following:

6.3.10.1 Agencies may transfer the web form(s) that initially connects the user to a database and/or server as the gateway for user input in either HTML or XML.

6.3.10.2 The business process that filters input to the database should be described in *Structured Query Language* (SQL) form (ANSI X3.135-1989[5] and ISO/IEC 9075) and may be written specifically for the transfer or, if available, extracted from software programs that connect the database with the web (e.g., *Common Gateway Interface* (CGI) scripts *or Java Database Connectivity* (JDBC) programs). Alternatively, agencies may provide a description of this business process as free text.

6.3.10.3 Databases must be transferred as described in 36 CFR 1228.270 (d)(1)[6].

6.4 Transfer Documentation

[5] ANSI Standards may be purchased through ANSI Headquarters (DC office) 1819 L. Street, NW, Sixth Floor, Washington, DC 20036 or ordered online at http://webstore.ansi.org.
[6] http://www.archives.gov/about_us/regulations/part_1228_l.html

Expanding Acceptable Management E-Gov Initiative
Transfer Instructions for Permanent Electronic Records

Transfer documentation is required for each transfer to enable processing and retrieval of the information. These requirements supplement the transfer documentation requirements in 36 CFR 1228.270(e). Agencies must submit a signed Standard Form 258, Agreement to Transfer Records to the National Archives of the United States (SF 258) for all transfers of web content records, as required by 36 CFR 1228.272.

For each transfer, if applicable, agencies must additionally supply documentation that identifies:

6.4.1 Web Platform and Server, including the specific software application(s) and, where available, intended browser application(s) and version(s).

6.4.2 Name of the web site.

6.4.3 URL of the web site, including the file name of the starting page of the transferred content.

6.4.4 Organizational unit primarily responsible for managing web site content.

6.4.5 Method and date of capture.

- If harvested, also include the application used with either a URL to the application's web site or a description of the harvester's capabilities and the log file(s) generated by the harvester that document the harvesting process.
- If PDF, also include the software and version used to capture the PDF.
- If manual, only method and date of capture are needed.
- If more than one method is used, clearly identify which content was captured by which method.

6.4.6 Contact information for individual(s) responsible for the capture.

6.4.7 The name and version of any content management system used to manage files on the web.

6.4.8 All file names, inclusive of both the path (or directory) name and the file name itself.

6.4.9 The business logic and web interfaces clearly identified with each corresponding database.

6.4.10 If available, a site map of the web site from which the web content records originated.

7.0 TRANSFERRING WEB CONTENT

7.1 An agency may use any of three methods to create the web content file to be transferred to NARA: (1) web harvesting, (2) capturing in Portable Document Format (PDF), or (3) manually downloading and copying. Appendix B summarizes these methods of capture and discusses the advantages and disadvantages of each. Appendix C provides basic guidance on how formats common to the web should be handled using each method.

7.1.1 If harvesting, make sure that the harvester settings support the terms and conditions of transfer articulated in the records schedule.

7.1.2 To ensure proper processing by NARA, the file name of each file transferred must be less than 100 characters and the file name with file path must be less than 255 characters. (Most web harvesters have the ability to truncate long file names while preserving functionality.)

7.1.3 Some file formats may not be captured when web harvesting or capturing content in PDF and should be acquired manually if considered part of the web content record. (See Appendix C)

7.1.4 Web content captured and transferred as PDF must meet both the transfer documentation requirements in section 6.4 of this guidance and the requirements in the NARA PDF Transfer Guidance issued March 31, 2004: Expanding Acceptable Transfer Requirements: Transfer Instructions for Permanent Electronic Records: RECORDS IN PORTABLE DOCUMENT FORMAT (PDF). [7]

7.2 Transfer Mechanisms

7.2.1 Agencies must transfer web content records using the approved and appropriately labeled media and methods listed in 36 CFR 1228.270(c).[8]

7.2.2 Web content records must not be compressed (e.g., WinZip, PKZIP) or aggregated (e.g., TAR) for purposes of transfer unless NARA has approved the transfer in compressed or aggregated form in advance. In such cases, NARA may require the agency to provide the software to decompress the records [see 36 CFR 1228.270(d)].

[7] http://www.archives.gov/records_management/initiatives/pdf_records.html
[8] http://www.archives.gov/about_us/regulations/part_1228_l.html.

8.0 CONTACT INFORMATION

For assistance in scheduling web content records, or to discuss how to handle permanent web content records that do not meet the specifications in this guidance, please contact your agency appraisal archivist in NARA's Life Cycle Management Division (NWML). The NWML general telephone number is 301-837-3560.

For assistance in transferring web content records to NARA, please contact the Electronic and Special Media Records Services Division (NWME), 8601 Adelphi Road, College Park, MD 20740. The NWME general telephone number is 301-837-3420.

Expanding Acceptable Management E-Gov Initiative
Transfer Instructions for Permanent Electronic Records

APPENDIX A
GLOSSARY

Application Program Interface (API). A specific method prescribed by a computer operating system or by an application program by which a programmer writing an application program can make requests of the operating system or another application.

Browser. A program used to view web content downloaded from the World Wide Web.

Client. The requesting program or user in a client/server relationship. For example, the user of a web browser is effectively making client requests for pages from servers all over the web. The browser itself is a client in its relationship with the computer that is getting and returning the requested HTML file.

Client/Server Architecture. The general structure of the web in which the load for processing web content is shared between a user's computer running a browser and a server computer that manages web content on a continuing basis and listens for web requests from clients to provide web content (they can be the same computer).

Common Gateway Interface (CGI). A standard way for a web server to pass a web user's request to an application program and to receive data back to forward to the user.

Document Type Definition (DTD). A specification that accompanies a SGML or XML document and identifies the content and relationships of markup elements in the document.

Domain name. That portion of the URL that precedes the domain extension (e.g., DomainName.gov). For example, while www.whitehouse.gov and www.whitehouse.gov/omb are both in the same site or domain, www.usda.gov and www.nal.usda.gov are not in the same site or domain in that one is registered as a primary domain and the other as a secondary domain.

Extensible Hypertext Markup Language (XHTML). A reformulation of HTML 4. When HTML is mentioned assume XHTML as well.

Extensible Markup Language (XML). A document markup language adopted by the W3C as a standard for the web that will eventually replace HTML. XML allows programmers to write their own markup language elements or "tags". XML separates the presentation, structure and content with style sheets schemas and content marked up with the "tags".

File fragment. Any part of a web document that is combined with the referenced HTML file through server or browser functionality. Although the term is explicitly associated with W3C's recommendation for File Fragment Interchange with XML, it is used here to also include concepts like Server Side Includes (SSI).

Hypertext Markup Language (HTML). The set of markup symbols or codes inserted in a file intended for display on a World Wide Web browser page. HTML is the language of today's web.

Hypertext Transfer Protocol (HTTP). The set of rules for transferring files (text, graphic images, sound, video, and other multimedia files) on the World Wide Web. As soon as a web user opens a web browser, the user is indirectly making use of HTTP. HTTP is an application protocol that runs on top of the foundation protocols for the Internet, the TCP/IP suite of protocols.

Java Database Connectivity (JDBC). An application program interface (API) specification for connecting programs written in Java to the data in databases.

Secure Hypertext Transfer Protocol (HTTPS). A security-enhanced version of HTTP providing a variety of mechanisms to enable confidentiality, authentication, and integrity. When HTTP is mentioned assume HTTPS as well.

Semantic Web. An extension of the web using Resource Description Format (RDF) in which content is made meaningful allowing sharing and reuse across application, enterprise, and community boundaries.

Server. A program that awaits and fulfills requests from client programs on the same or other computers. On the web, servers are the location of most web content. The server may be a complex system of multiple tiers of applications that all interact via configurations that are specific to that computer.

Server Side Includes (SSI). A file fragment or part of a document that a server can include in an HTML file before it sends it to the browser.

Structured Query Language (SQL). Both an ANSI and an ISO standard, SQL is recognized as the most portable way of requesting information from relational databases, and is used as one of the main access routes for all such databases, and for certain classes of object-oriented databases.

Style sheets. A generic term to describe how documents are presented on screens or in print. Cascading Style Sheets (CSS) is a specific type of style sheet used for HTML and XML. eXtensible Stylesheet Language Transformations (XSLT) is another type of style sheet used specifically with XML.

Uniform Resource Identifier (URI). A compact string of characters for identifying an abstract or physical resource. The most common form of URI is the web page address, which is a particular form or subset of URI called a URL.

Uniform Resource Locator (URL). The unique address for a file that is accessible on the web.

Web Content. Information that is sent from a server to a browser via Hypertext Transfer Protocol (HTTP) when a URL has been activated.

Web Content Record: Information that meets the definition of a Federal record and is provided via an agency's web site.

Web Services. The programmatic interfaces that allow application to application communication on the web.

Web Site. A related collection of web content identified by a domain name .

Wireless Web. The use of the web through a wireless device, such as a cellular telephone or personal digital assistant (PDA).

World Wide Web Consortium (W3C). An organization that "…develops interoperable technologies (specifications, guidelines, software, and tools) to lead the web to its full potential…" and is "…a forum for information, commerce, communication, and collective understanding." (http://www.w3.org)

XML. See *Extensible Markup Language*.

XML Schema. A convention to define the structure, content and semantics of XML documents. Schemas are replacing DTDs for XML.

Expanding Acceptable Management E-Gov Initiative - Transfer Instructions for Permanent Electronic Records
Appendix B – Methods of Capture

This appendix identifies the pros and cons of the three methods of capturing web content records. It should be used with Appendix C to determine the most appropriate capture method.

Method of Web Capture	Pros	Cons	Requirements for use in transfer
Local Harvesting There are a number of "web harvesters" or "web crawlers" available, both proprietary and open source. These harvesters, once given the URL of a web site homepage, will download or copy the designated content of a web site onto a local medium. These programs have varying amounts of success. Some format types may not be downloaded. (See Appendix C)	• Many "web harvesters" and "web crawlers" available. • Automatic download of desired web content onto local media. • Many automatically redirect internal links. • Some provide facility to add metadata (e.g., comment external links) • Dynamically rendered content (i.e., Cold Fusion (.cfm) files, PHP (.php) files and Active Server Page (.asp) files) are automatically converted to browser readable HTML. • Automatically truncates file name and file path size to acceptable limits.	• Evolving technology • Capture of client-side functionality depends on programming language used (e.g. VBScript may not be captured) • Server-side functionality not captured	Agencies should include comments or external documentation describing the content of external links that they consider significant to the content of the web site being transferred. Comments should include the title of linked content, the source agency or institution, and the significance of the content to the web content records being transferred.
PDF Web Capture is a PDF 1.3 feature. Acrobat 4 and later viewers will capture HTML, PDF, GIF, JPEG, and ASCII text files as PDF files.	• Automatic download of desired web content onto local media. • Automatically redirects internal links. • Represents external links as URLs	• All rendered content is in the form of static PDF images, except that links display and work. Other functionalities of the pages as they were on the web, e.g., image captions, will be lost. • Not appropriate complex web sites or those with large	• Web content captured and transferred as PDF must meet both the transfer documentation requirements in section 6.4 of this guidance and the requirements in the NARA PDF Transfer Guidance issued March 31, 2004: Expanding Acceptable Transfer Requirements: Transfer

Method of Web Capture	Pros	Cons	Requirements for use in transfer
		amounts of web content.	Instructions for Permanent Electronic Records: RECORDS IN PORTABLE DOCUMENT FORMAT (PDF). • Agencies should include comments or external documentation describing the content of external links that they consider significant to the content of the web site being transferred. Comments should include the title of linked content, the source agency or institution, and the significance of the content to the web content records being transferred.
Manual Download and copy web content records directly from the server to a transfer medium. Note that the contents of a web page often may not all be located in the same subdirectory.	• Download and copy web content records directly from a web server to a transfer medium. • Not dependent on capture software	• Difficulty in identifying and locating desired web content (e.g., files are not stored in the same directory or location). • Dynamically rendered content (i.e. Cold Fusion (.cfm) files, PHP (.php) files, Server Side Inclusions (SSI) and Active Server Page (.asp) files) may not be easily interpreted without server-side software.	• Agencies must redirect links that are pointing to content within the transfer. • Agencies should include comments or external documentation describing the content of external links that they consider significant to the content of the web site being transferred. Comments should include the title of linked content, the source agency or institution, and the significance of the content to the web content records being transferred. • Agencies must convert dynamically rendered content (i.e.

Method of Web Capture	Pros	Cons	Requirements for use in transfer
			Cold Fusion (.cfm) files, PHP (.php) files, Server Side Inclusions (SSI) and Active Server Page (.asp) files) to an HTML version that can be read by a browser.

Expanding Acceptable Management E-Gov Initiative - Transfer Instructions for Permanent Electronic Records
Appendix C – Guide to appropriate capture methods for specific web content record formats

Content Format(s)	Harvest	PDF Capture	Manual
HTML, XHTML	As harvested	As captured	Must be accompanied by Cascading Style Sheets (CSS) if used
XML	As harvested	As captured	Must be accompanied by schema or DTD and style sheet(s)
ASP, PHP, Cold Fusion, SSI	As harvested	As captured	Must convert dynamically rendered content (i.e. Cold Fusion (.cfm) files, PHP (.php) files, Server Side Inclusions (SSI) and Active Server Page (.asp) files) to an HTML version that can be read by a browser.
Images as part of HTML file (also may apply to audio and video)	As harvested. If not harvested, audio and video must be collected manually	As captured. If not captured, audio and video must be collected manually	Links from HTML must function as transferred
Stand alone audio, video, images, PDF Word Processing, plain text and RTF or other proprietary formats	Limited ability to harvest	Limited ability to capture	Must be transferred as binary files (binary large objects) from server
Database	Will not harvest, must be manually captured	Will not capture as PDF, must be manually captured	Must be transferred as per 36 CFR 1228.270 (d)(1). Include web interface page and business connection as an SQL statement or in free text
GIS	Will not harvest, must be manually captured	Will not capture as PDF, must be manually captured	Must be transferred as per Expanding Acceptable Transfer Requirements: Transfer Instructions for Permanent Electronic Records: DIGITAL GEOSPATIAL DATA RECORDS [9]

[9] http://www.archives.gov/records_management/initiatives/digital_geospatial_data_records.html